Towards a Slovene Dependency Treebank

Sašo Džeroski,[*] Tomaž Erjavec,[*] Nina Ledinek,[†]
Petr Pajas,[‡] Zdenek Žabokrtsky,[‡] Andreja Žele[♠]

* Department of Knowledge Technologies, Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
saso.dzeroski@ijs.si, tomaz.erjavec@ijs.si
† Šoštanj, Slovenia
nina.ledinek@siol.net
‡ Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25, CZ-118 00 Prague, Czech Republic
pajas@ufal.ms.mff.cuni.cz, zabokrtsky@ckl.mff.cuni.cz[1]
♠The Fran Ramovš Institute for the Slovenian Language
Scientific and Research Centre of the Slovene Academy of Sciences and Arts
Novi trg 4, SI-1000 Ljubljana, Slovenia
andreja.zele@zrc-sazu.si

Abstract

The paper presents the initial release of the Slovene Dependency Treebank, currently containing 2000 sentences or 30.000 words. Our approach to annotation is based on the Prague Dependency Treebank, which serves as an excellent model due to the similarity of the languages, the existence of a detailed annotation guide and an annotation editor. The initial treebank contains a portion of the MULTEXT-East parallel word-level annotated corpus, namely the first part of the Slovene translation of Orwell's "1984". This corpus was first parsed automatically, to arrive at the initial analytic level dependency trees. These were then hand corrected using the tree editor TrEd; simultaneously, the Czech annotation manual was modified for Slovene. The current version is available in XML/TEI, as well as derived formats, and has been used in a comparative evaluation using the MALT parser, and as one of the languages present in the CoNLL-X shared task on dependency parsing. The paper also discusses further work, in the first instance the composition of the corpus to be annotated next.

1. Introduction

Syntactically annotated corpora, treebanks, are an important language resource: on the one hand they facilitate the study of theoretical syntax and the syntax of particular languages as evidenced on real, naturally occurring and discourse-connected sentences; on the other, they serve as testing and, increasingly, training datasets for syntactic parsers, useful components of applications involving the processing of natural languages. With the recent upsurge in research on statistical parser induction, especially the latter holds great promise.

Slovene (or Slovenian) is a South-Slavic language, spoken predominantly in Slovenia, with about two million speakers. As other Slavic languages it is characterized by rich inflection (it is e.g. one of the very few languages that morphologically distinguishes the dual number), and free word order, however with complex constraints regarding e.g. clitic placement and topic-focus prominence. While a number of traditional grammars have been written for Slovene, the most comprehensive being Toporišič (1984), and various aspects of the language have been studied in the generative framework, there has so far been no wide-coverage formal (i.e. computational) grammar written for the language, and neither have there been any previous attempts to produce a Slovene treebank. The state-of-the art of Slovene written language resources so far extended to producing and making available monolingual and parallel morphosyntactically tagged and lemmatised corpora, e.g. (Erjavec et al., 1998, Erjavec, 2002).

The Slovene Dependency Treebank (SDT) is conceived as a long term project, with, currently, no dedicated funding. It started already in 2003, when the initial decisions regarding the formalism and software platform were made and implemented, and the automatically annotated corpus was compiled. Since then, the effort has gone into manually annotating the first 2,000 sentences / 30,000 words with analytic tree structures and preparing the annotation manual. This first phase has just recently been finished, and the results already used in a few experiments. This paper serves to document this first stage of building the Slovene Dependency Treebank and set out our plans for the future.

While not many sentences have so far been annotated, we have created the essential infrastructure for the undertaking: Section 2 introduces our theoretical and implementation model, Section 3 explains how we operationalised and adapted it to Slovene, Section 4 gives two experiments that SDT has participated in, Section 5 discusses our future plans and dilemmas in how to extend the corpus, while Section 6 gives some conclusions.

2. The Prague Dependency Treebank

The first task that needs to be undertaken in building a treebank for a new language is developing or adopting a theoretical and practical framework in which to annotate the texts. For Slovene, we were lucky in that we had a very good model in the Prague Dependency Treebank (LDC, 2001).

[1] Supported by grant 1ET101120503.

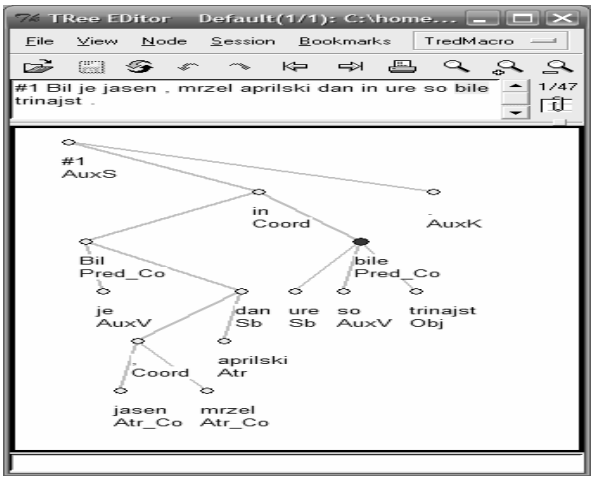

Figure 1 Annotated sentence form SDT in TrEd

The most important aspect of PDT as regards SDT is that the Czech language is much closer to Slovene than are English or any other Germanic or Romance language, i.e. languages for which most treebanks and associated methodology already exist. Therefore not only the theoretical model (dependency grammar, in particular Functional Generative Grammar, for PDT) but also the concrete solutions for Czech are immediately relevant to Slovene. And although there exist other Slavic treebanks, e.g. Bulgarian (Simov et al., 2002) and Russian (Boguslavsky et al., 2000), the PDT project has also made available their annotation manual and an editor for manual annotation and visualisation of the annotated corpus.

The PDT distinguishes two levels of syntactic annotation. The analytic level encodes shallow syntactic dependencies, which are assigned to all (and only) tokens appearing in the sentences, while the tectogrammatical level encodes deeper, already semantic relations between the constituents of the sentence. For Slovene we are currently concentrating only on the analytic level.

2.1. The PDT annotation manual

In order to annotate a treebank, esp. by a group of annotators, it is essential to have a comprehensive annotation manual, which explains the annotation conventions and covers the grammar of the language, illustrating it with dependency fragments and examples. Producing such a manual is extremely difficult and time consuming, and has to proceed hand in hand with the annotation, in an interactive process of analysis, description, and refinement. For the PDT, a 300+ page (English) manual, covering the annotations at the analytical level and containing instructions for annotators, was produced (Hajič et al., 1997) and made publicly available. As Slovene is grammatically relatively close to Czech, we started off using the PDT as is, but keeping track and commenting on the clauses where Slovene grammar differs from Czech. Currently, the most detailed analysis and recommendations were produced for the treatment of the most complex dependency type and part-of-speech, namely predicate and verb (Ledinek, 2005).

At the end of the first stage of annotation, we started substituting the examples in the Czech manual with ones in Slovene, and modifying the differing clauses to reflect the behaviour of Slovene. This process is still far from finished, as the corpus data keep presenting us with new challenges. We therefore consider the development of the SDT annotation manual as an on-going process, likely to last for some time into the future.

2.2. The TrEd annotation editor

A vital component of a treebanking project is an annotation editor that enables the visualisation and hand-editing of dependency structures. A good editor can make the life of annotators much more bearable, as well as lowering the error rate and speeding up production. In the scope of PDT, a powerful dedicated editor TrEd (Hajič et al., 2001) was produced, and, again, made publicly available. The editor is written in Perl/Tk, and reasonably portable across platforms. It enables navigation through files and sentences, drag-and-drop structure annotation and pick lists for dependency labels. It is highly configurable, and provides several input and output formats, e.g. XML/TEI and GIFs of the annotated trees. An example screenshot of TrEd, showing the first sentence in the novel, is given in Figure 1.

3. The initial Treebank

The process of corpus annotation proceeded by first converting the source XML files (Section 3.1) into TrEd format and splitting them into chunks of approx. 50 sentences each.[2] These files were then annotated, first with a parser (Section 3.2.) and these annotations manually corrected. Finally, the dependency annotations in the TrEd files were merged with the original data, and the resulting corpus made into an XML document (Section 3.3).

3.1. MULTEXT-East Orwell's 1984

For the first sample to be annotated we considered it more important that the text was pre-processed as much as possible, rather than aiming at a (in any way) representative sample of the language. We therefore chose a part of the MULTEXT-East parallel corpus (Erjavec, 2004), which is encoded in XML according to the Text Encoding Initiative Guidelines P4 (Sperberg-McQueen and Burnard, 2002), is sentence aligned with English and a number of other translations, and is tagged with hand-validated context-disambiguated lemmas and morphosyntactic descriptions, which follow the EAGLES and MULTEXT guidelines. This choice not only leads to a richer and more interchangeable data resource, but also enables initial dependency structures to be put in place automatically, as discussed below.

The text that has been so far annotated is Part one (i.e. the first third) of the novel "1984" by G. Orwell. While choosing this text does have a number of technical advantages, it also has its drawbacks: we were annotating only a single text, which is translated rather than original Slovene, and not a strikingly good translation at that. As it is fiction, it also contains long sentences, with lots of direct speech, making it hard to annotate. Finally, it contains "anomalous" language (Newspeak) which led to considerable problems in the analysis; or, depending on the viewpoint, makes it even more interesting.

[2] The size of 50 sentences per file was chosen as this is approximately what a person can annotate in a day; this sets a simple yardstick of progress for the annotators.

```
<text id="Osl." lang="sl">
 <body>
  <div type="part" id="Osl.1">
   <div type="chapter" id="Osl.1.2">
    <p id="Osl.1.2.2">
     <s id="Osl.1.2.2.1">
      <w id="s1t1" afun="Pred" parallel="Co" dep="s1t8" lemma="biti" ana="Vcps-sma">Bil</w>
      <w id="s1t2" afun="AuxV" dep="s1t1" lemma="biti" ana="Vcip3s--n">je</w>
      <w id="s1t3" afun="Atr" parallel="Co" dep="s1t4" lemma="jasen" ana="Afpmsnn">jasen</w>
      <c id="s1t4" afun="Coord" dep="s1t7">,</c>
```

Figure 2. SDT in the canonical TEI format

3.2. A rule-based parser for Slovene

To make the work of manual annotation easier we built a parser for Slovene, which assigns initial analytic tree structures to the corpus. The parser takes advantage of the gold-standard morphosyntactic annotations already present in the corpus, which give quite detailed word-level syntactic information. As opposed to parsing the word-forms directly, this approach obviates the need for a lexicon, and significantly reduces the input ambiguity.

The parser is written in Perl, and implements bottom-up reduction-based parsing, trying to apply its rules (hard-wired into the code), inside a sliding window over the (partially processed) input. It also uses fall-back rules and rules for top-down parsing, which take care of dependencies outside the scope of the sliding window. The parser is quite fast, as it does no backtracking; if the parser makes a bad decision which is only detected later, it tries to correct the error by applying post-processing rules. The parser correctly annotated about 60% of the dependencies in "1984", significantly speeding up the manual annotation.

3.3. Corpus size and encoding

The prototype release of the Slovene Dependency Treebank contains 1984 sentences or just over 30,000 tokens, manually annotated with analytic-level dependency trees, as well as word-level morphosyntactic descriptions and lemmas. The SDT is available in several formats, with the canonical being the one from the MULTEXT-East corpus, i.e. TEI P4, with extensions to token attributes, which encode a pointer to the parent node and the dependency.

The complete Slovene Dependency Treebank is composed of the TEI corpus header, and three TEI documents, each containing its own header, followed by the body. The first document contains the formal MULTEXT-East morphosyntactic specifications that give the feature decomposition of the word-level morphosyntactic descriptions used in the corpus text. The second gives the feature library for the analytical functions, which manifest themselves in two attributes on tokens. The third document contains the currently single component of the corpus, namely the first part of the Slovene translation of the MULTEXT-East "1984" novel, with the tokens additionally annotated with dependency relations and their functions. An example sentence from the corpus is given in Figure 2

4. Current experiments on SDT

The current interest in inductive parsing is well brought out by the interest in SDT, despite its small size.

The section describes two experiments: the first concerns using the MALT parser in a comparative evaluation of parsing accuracy, and the second the CoNLL-X shared task on dependency parsing.

4.1. Parsing with MALT

The SDT was first used in the context of a study on parsing accuracy of (mainly) Italian (Chanev, 2005), using the MALT parser (Nivre & Hall, 2005); Slovene scored significantly lower (by about 10% of absolute accuracy) than Italian, although this was probably largely due to the smaller size of the Slovene dataset. The accuracy reported in the paper was also obtained on a previous version of the SDT, which was smaller, and contained more inconsistencies.

The experiment was later re-run[3] on the current version of SDT (in the CoNLL format), while also using the so called m7 MALT learning model, instead of the previous m3 and m4. The results in this improved configuration are by around 5% better, and are, using 10-fold cross validations on contiguous stretches of text, and taking the gold standard morphosyntactic tags, 63.42% for labelled and 74.20% for unlabelled precision. As the learning curve at this size of corpus is relatively steep, this gives us hope that it will be rather soon possible to induce a useful parser from the treebank.

4.2. The CoNLL-X shared task

The second "experiment" concerns the inclusion of SDT into the shared task on dependency parsing organised in the scope of the 10th Conference on Computational Natural Language Learning, CoNLL-X.[4] At the time of this writing it is too early to say what the results were or even if the Slovene dataset was chosen by any of the contestants. In any case, even the preparation of the data into the shared task format was useful. Namely, the data format for CoNLL-X, as well as that for MALT makes certain assumptions about the input which are not met by SDT; the most important is that in SDT (or PDT) punctuation is allowed to act as an internal node in the tree (i.e. is allowed to be a head), while in CoNLL-X this is not allowed. Similarly, MALT discards the punctuation, leading to problems where a punctuation token has children. Another difference is that in PDT the root of the sentence is a virtual node, and more than one token can be its child; in CoNLL-X/MALT the root must be an actual token.

[3] Atanas Chanev, personal communication, December 20th, 2005.
[4] CoNLL-X conference: http://www.cnts.ua.ac.be/conll/, Shared task on dependency parsing: http://nextens.uvt.nl/~conll/

We therefore wrote an XSLT script that converts the TEI PDT format to that of CoNLL-X. Which conversions to perform, e.g. demoting punctuation, is set by parameters to the script. The script performs the required conversions and formats the data into the CoNLL-X-specified tabular file, where, inter alia, the morphosyntactic codes are expanded from their short form into full attribute-value pairs. In addition to the canonical XML/TEI encoding, the SDT is now thus also available in a widely recognised format suitable for feeding to inductive dependency parsers.

5. Extending the corpus

Further work on SDT will proceed in two directions. We will give priority to finishing the annotation manual for Slovene and extending the pool of annotators. Simultaneously, we need to prepare the next part of the corpus for annotation. Here, the immediate question is which texts to choose for such a corpus.

One relevant factor is the Penn Treebank, still the most popular syntactically annotated corpus. Compiling a Penn-comparable corpus for the SDT makes comparative studies and experiments easier. Furthermore, an exciting research opportunity has opened with the publication of the Prague Czech-English Dependency Treebank (Čmejrek et al., 2004; LDC, 2004), containing a translation of a portion of the Penn Treebank into Czech, with dependency annotations assigned to both languages. Therefore translating a portion of Penn into Slovene would bring with it the advantage of having a three-way parallel dependency annotated corpus, resulting in an ideal resource for MT and cross-lingual research. An especially interesting option is pursuing research into induction of grammar transfer rules between languages (Kuhn, 2004), attempting to learn the Slovene annotations from the Czech ones.

We would also like to make the treebank useful as soon as possible, which means choosing text types that are likely to be similar to those that will be parsed in potential applications. These are, however, likely to be Web texts, which makes this condition somewhat at odds with the previous one. However, if this option is chosen, the texts will most likely be taken from the FIDA+ corpus,[5] which is already annotated with lemmas and morphosyntactic descriptions.

6. Conclusions

The paper has presented the proto-release of the Slovene Dependency Treebank, which is modelled after the Prague Dependency Treebank. While the corpus is small, it has already been put to use in several experiments. In order to make the corpus maximally useful, we have packaged it in three formats, one TEI P4, one as the native TrEd format (.fs), and the third as a tabular file in the CoNLL-X shared task format. The SDT is freely available for research use; how to obtain it is explained on the project homepage at http://nl.ijs.si/sdt/.

We presented also our further work, to an extent contingent on funding, but basically concentrating on further adaptation of the manual for analytic annotation of the PDT to Slovene, and extending the annotated corpus with new material. Furthermore, we plan to start experiments in automatic parser induction.

[5]The FIDA+ corpus, http://www.fidaplus.net/, is the continuation of the FIDA project, http://www.fida.net/, which produced a reference Slovene corpus of a 100 million words.

References

Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., & Frid, N. (2000). Treebank for Russian: Concept, tools, types of information. *COLING-2000.*.

Chanev, A. (2005). Portability of Dependency Parsing Algorithms - an Application for Italian. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT'05).* Barcelona.

Čmejrek, M., Cuřín, J. Havelka, J., Hajič, J., Kuboň, V. (2004). Prague Czech-English Dependency Treebank: Syntactically Annotated Resources for Machine Translation. *LREC'04*.

Erjavec, T. Gorjanc, V., Stabej, M. (1998). Korpus FIDA. (The FIDA corpus). Conference *Jezikovne tehnologije za slovenski jezik.* Ljubljana: Jožef Stefan Institute.

Erjavec, T. (2006). The English-Slovene ACQUIS corpus. *LREC'06*.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *LREC'04*, pp. 1535-1538,.

Hajič J., Pajas P. and Vidová Hladká, B. (2001). The Prague Dependency Treebank: Annotation Structure and Support. *IRCS Workshop on Linguistic databases*, 2001, pp. 105-114.

Hajič J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A. (1997) *A Manual for Analytic Layer Tagging of the Prague Dependency Treebank.* UFAL Technical Report TR-1997-03, Charles University, Czech Republic.

Kuhn, J. (2004). Experiments in Parallel-Text Based Grammar Induction. In *Proc. ACL'04.*

Ledinek, N. (2005) Površinskoskladenjsko označevanje korpusa Slovene Dependency Treebank (s poudarkom na predikatu). *(Surface syntactic annotation of the Slovene Dependency Treebank (with focus on the predicate).* ***B.A.*** thesis. University of Ljubljana.

Linguistic Data Consortium. (2001). *Prague Dependency Treebank 1.* LDC2001T10.

Linguistic Data Consortium. (2004). *Prague Czech-English Dependency Treebank Version 1.0,* LDC2004T25.

Marcus, M. Beatrice, P. S. & Markiewicz, M.A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics,* 19/2.

Nivre, J., Hall, J. (2005). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. In *Proc. the Fourth Workshop on Treebanks and Linguistic Theories (TLT'05).* Barcelona.

Simov, K., Osenova, P., Slavcheva, M., Kolkovska, S., Balabanova, E., Doikoff, D., Ivanova, K., Simov, A., & Kouylekov, M. (2002). Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank *LREC'02*.

Sperberg-McQueen, C. M. and Burnard, L. (eds.) (2002). *Guidelines for Electronic Text Encoding and Interchange, the XML Version of the TEI Guidelines.* The TEI Consortium.

Toporišič, J. (1984). Slovenska slovnica (*Slovene Grammar*), Obzorja, Maribor.